CEP of On-line Discussions: Multilingual Text Analysis for Early Warning

A Perspective from Sandia's "Networks Grand Challenge"

An Address to the National Academies Standing Committee
for Technology Insight—Gauge Evaluate and Review (TIGER)

Philip Kegelmeyer, wpk@sandia.gov, `csmr.ca.sandia.gov/~wpk`

National Nuclear Security Administration

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

April 27, 2011

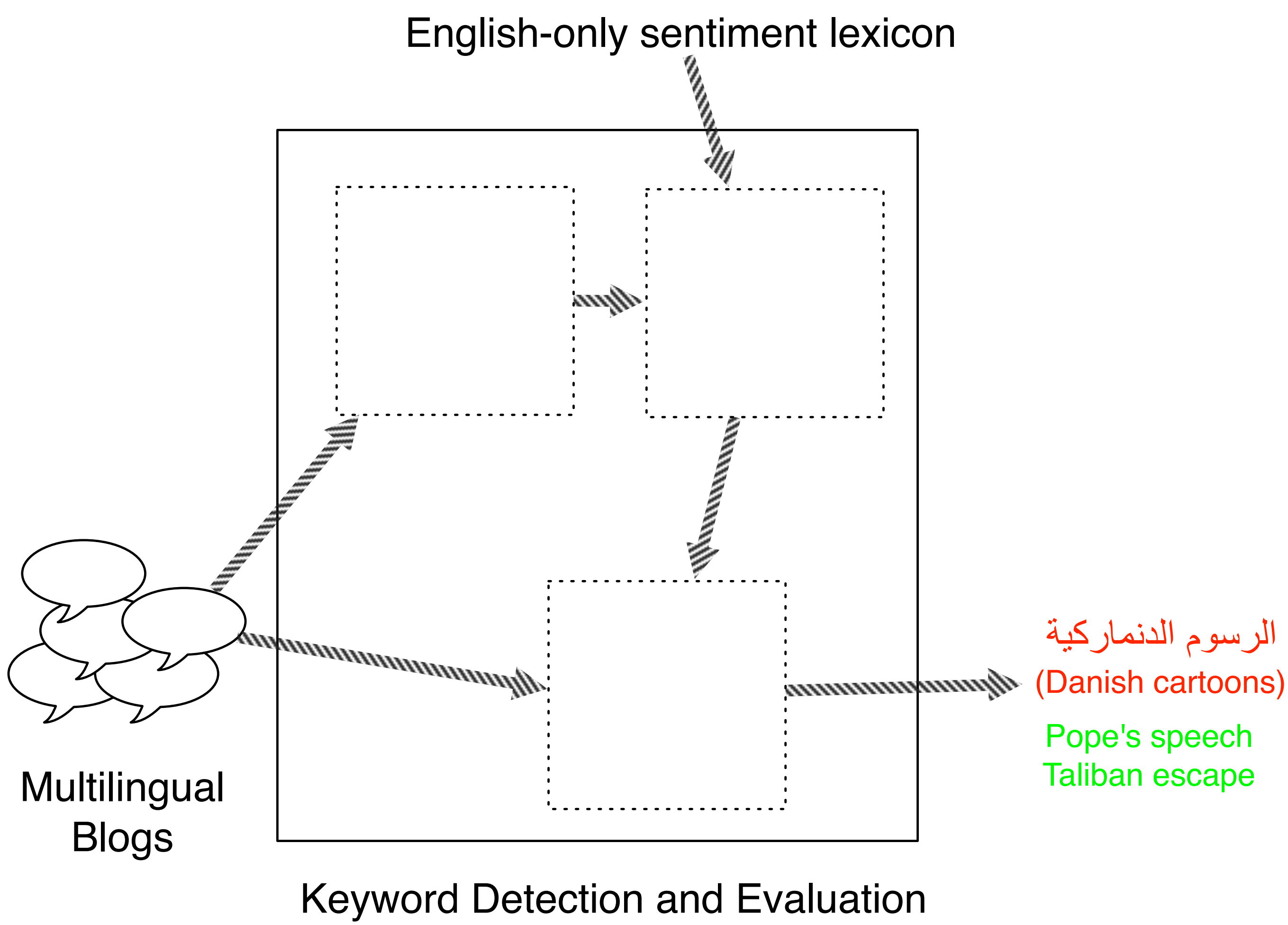
CEP: Questions To Address

- **What is our definition of CEP?**

  Daily monitoring and analysis of foreign language web blogs to find keywords of IC interest **and** to do predictive analysis as to whether a given keyword discussion will spill over into real world consequences.

  - "organization": everyone involved in a on-line discussion domain
  - "event": blog post or comment
  - "meaningful events": topically coherent blog posts
  - "analyzing their impact": forewarning of real world consequences
- How do we do the processing?
- What kind of results do we get?
- Who is using our service and for what?
- What are the computing and data specifications, limitations, metrics?
- What examples of the whole process and success stories?

The Early Warning Black Box

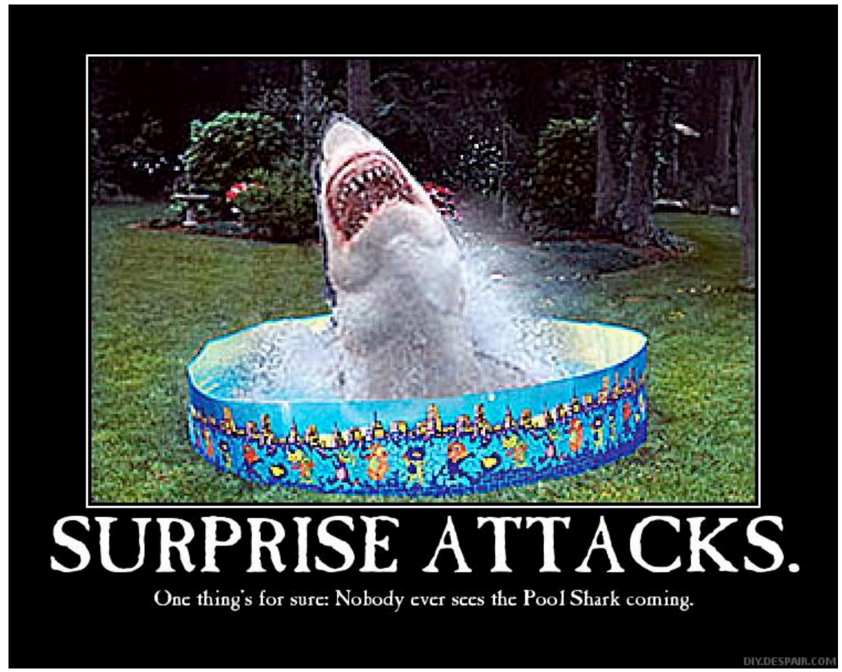


The Goal of the Networks Grand Challenge

"In this project, we build upon considerable existing Sandia capabilities in **scalable computing** and **advanced analysis algorithms**. We will understand and elicit the **needs of the intelligence community**, do basic research on **uncertainty** in the intelligence domain, research and evaluate **novel analysis algorithms**, and **implement** that research to address those needs to create a flexible, **interactive** capability for intelligence analysis on large datasets."

Two Problem Domains of Interest

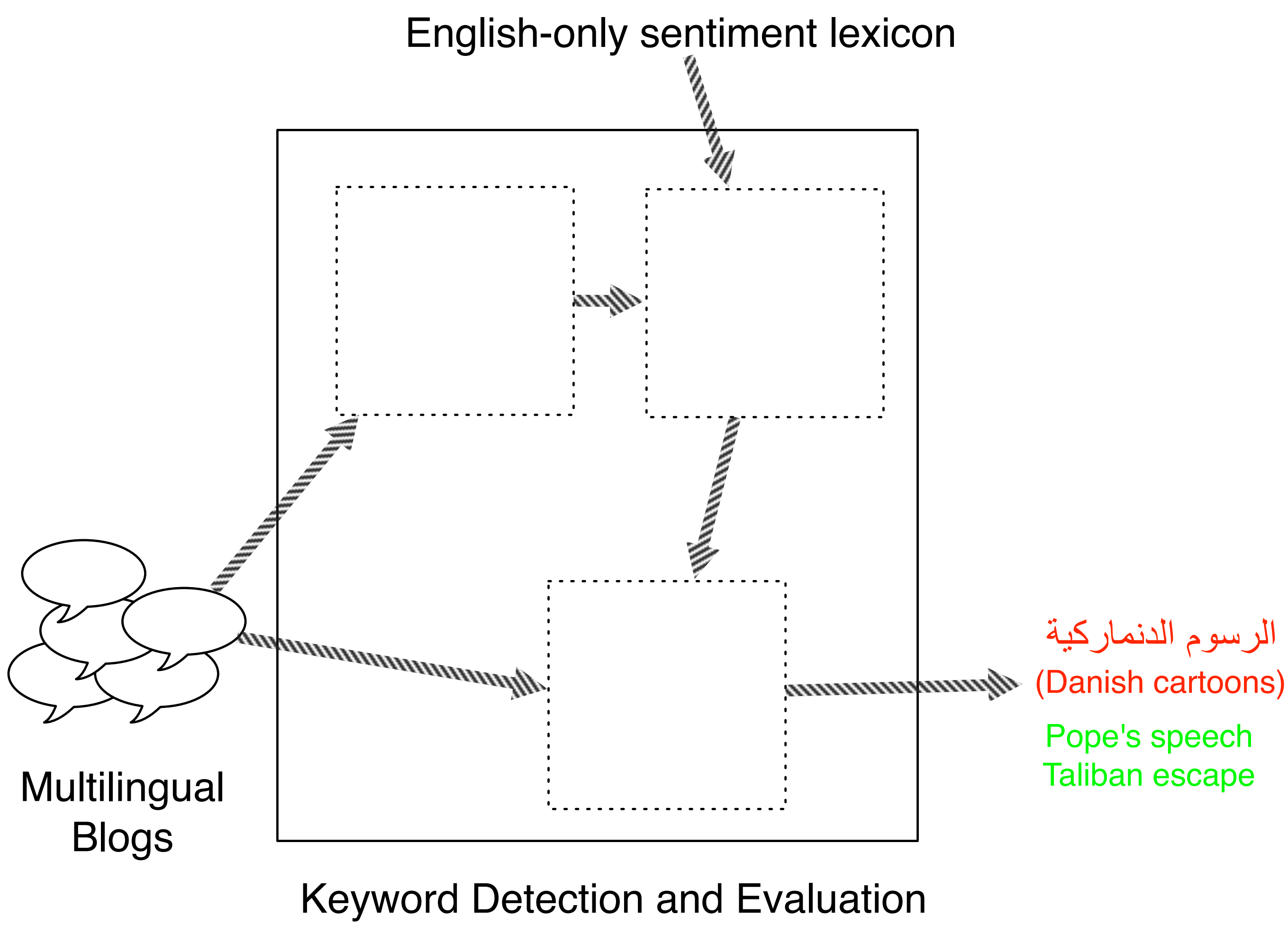
Cybersecurity[1]



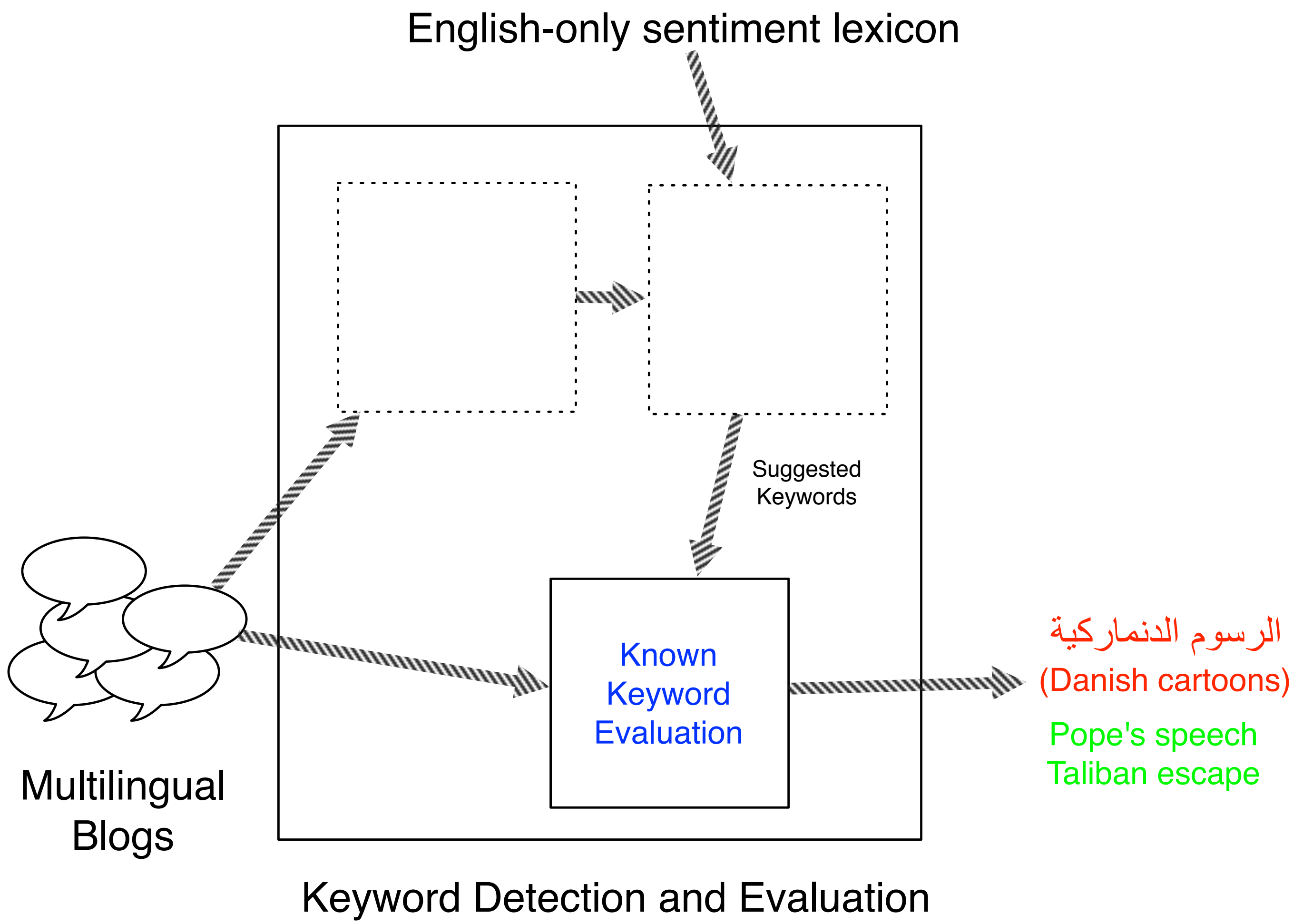
Technology Surprise

[1] *Economist*, Volume 391, Number 8630, May 9th, 2009

The Early Warning Black Box



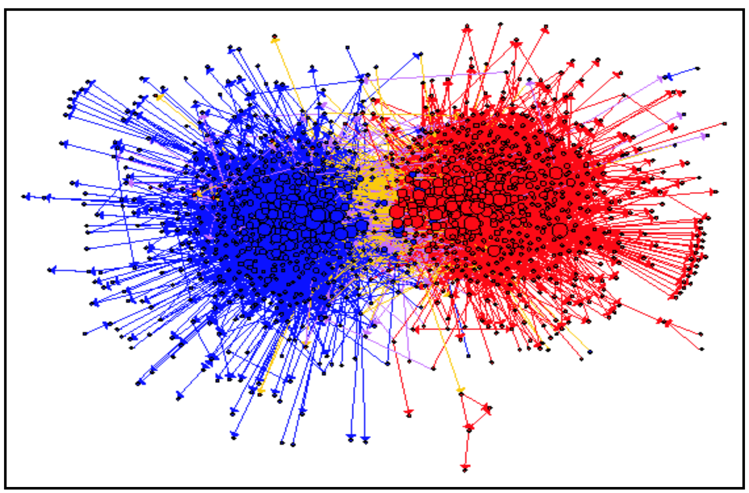
Evaluating Known Keywords

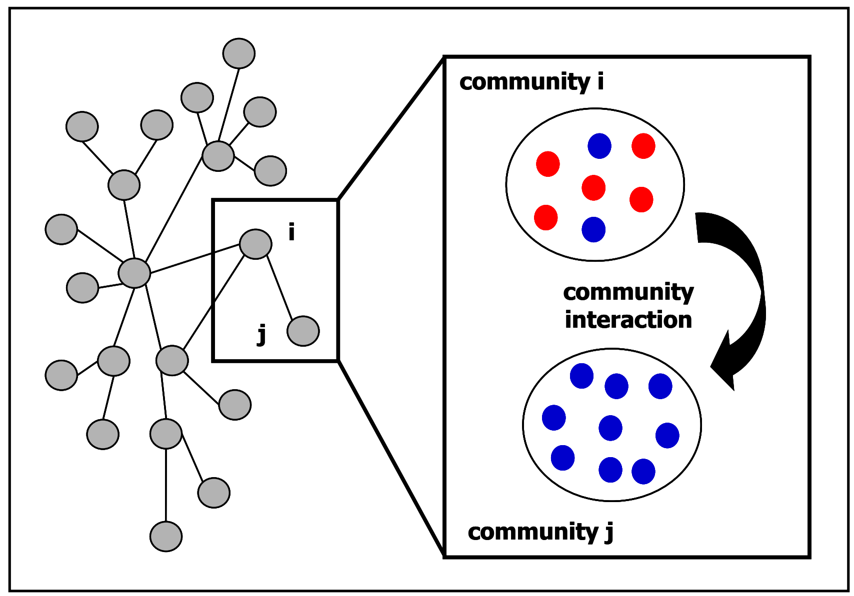


Multi-scale social dynamics model

A broad range of social dynamics phenomena can be usefully represented within a multi-scale modeling framework:

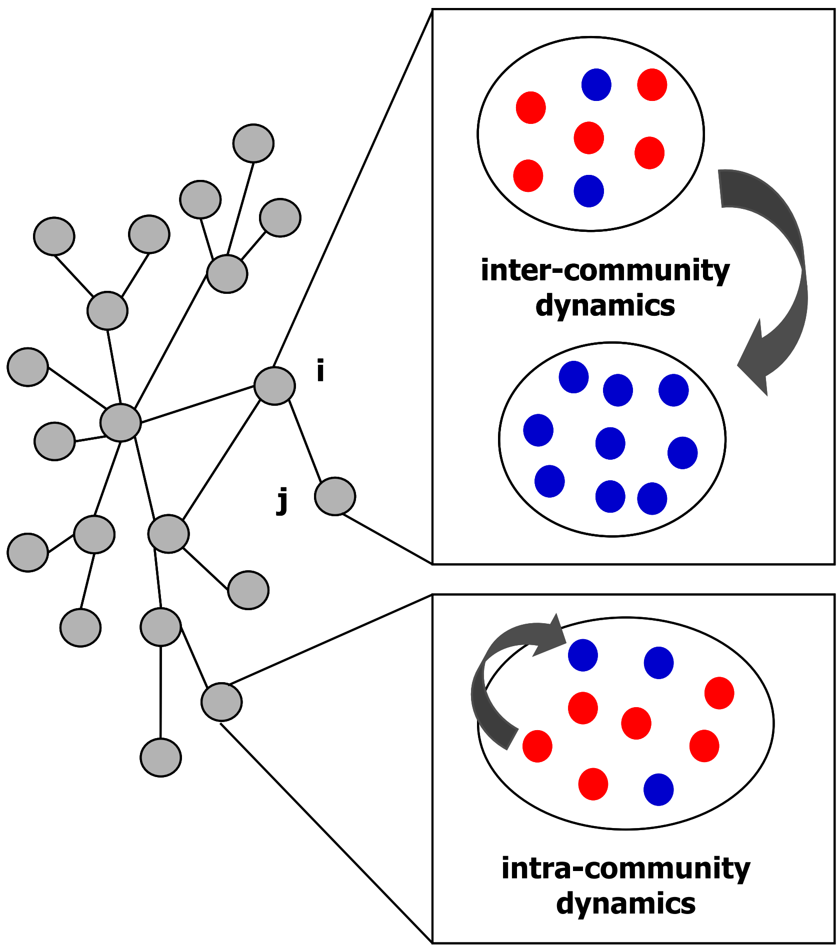
- micro-scale – behavior of individuals;
- meso-scale – interactions *within* social network communities;
- macro-scale – interactions *between* communities.



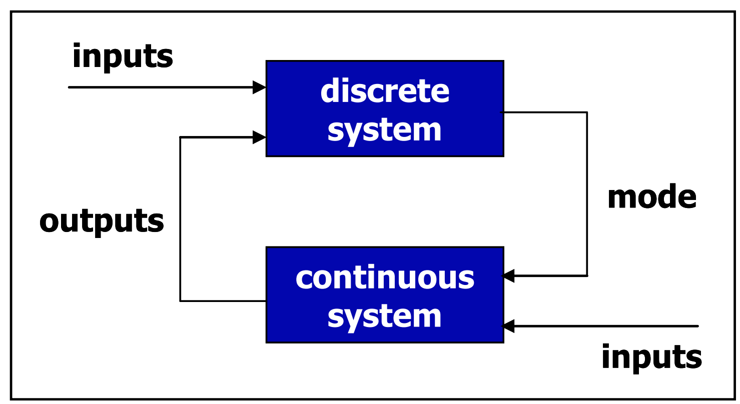


Stochastic Hybrid Dynamic Systems

Multi-scale social dynamics model: S-HDS realization



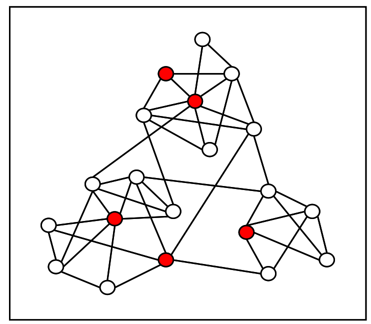
The stochastic hybrid dynamical system (S-HDS) model provides a natural and powerful formalism within which to represent multi-scale social dynamics (expressive, scalable, amenable to formal analysis).

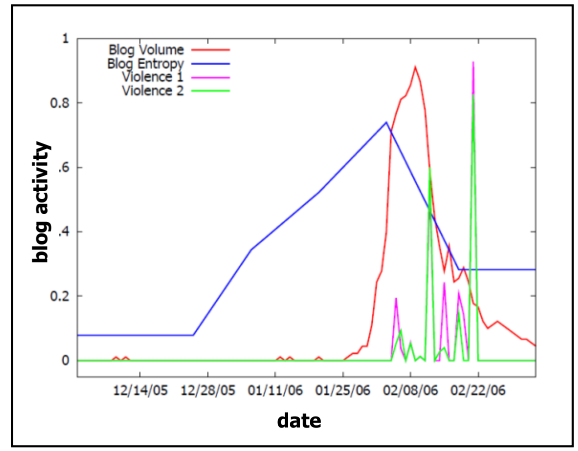


Entropy as an early indicator ...

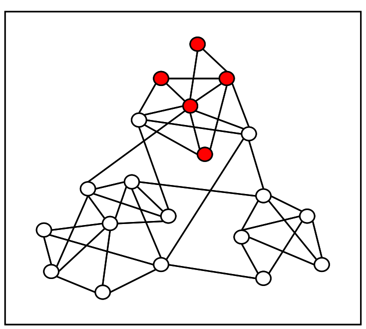
Blog post dispersion across communities is an useful early indicator of large mobilizations.

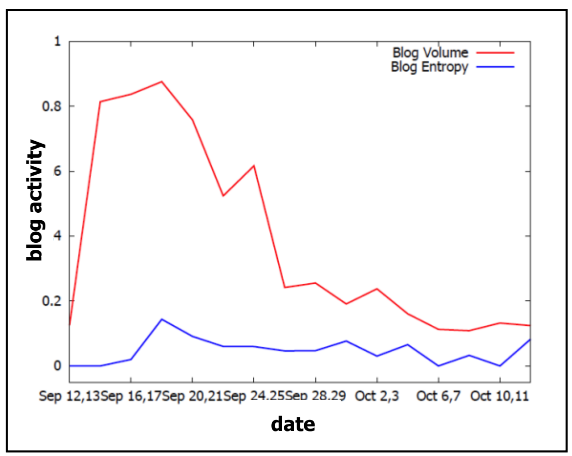
Danish Cartoons 1





Pope Lecture





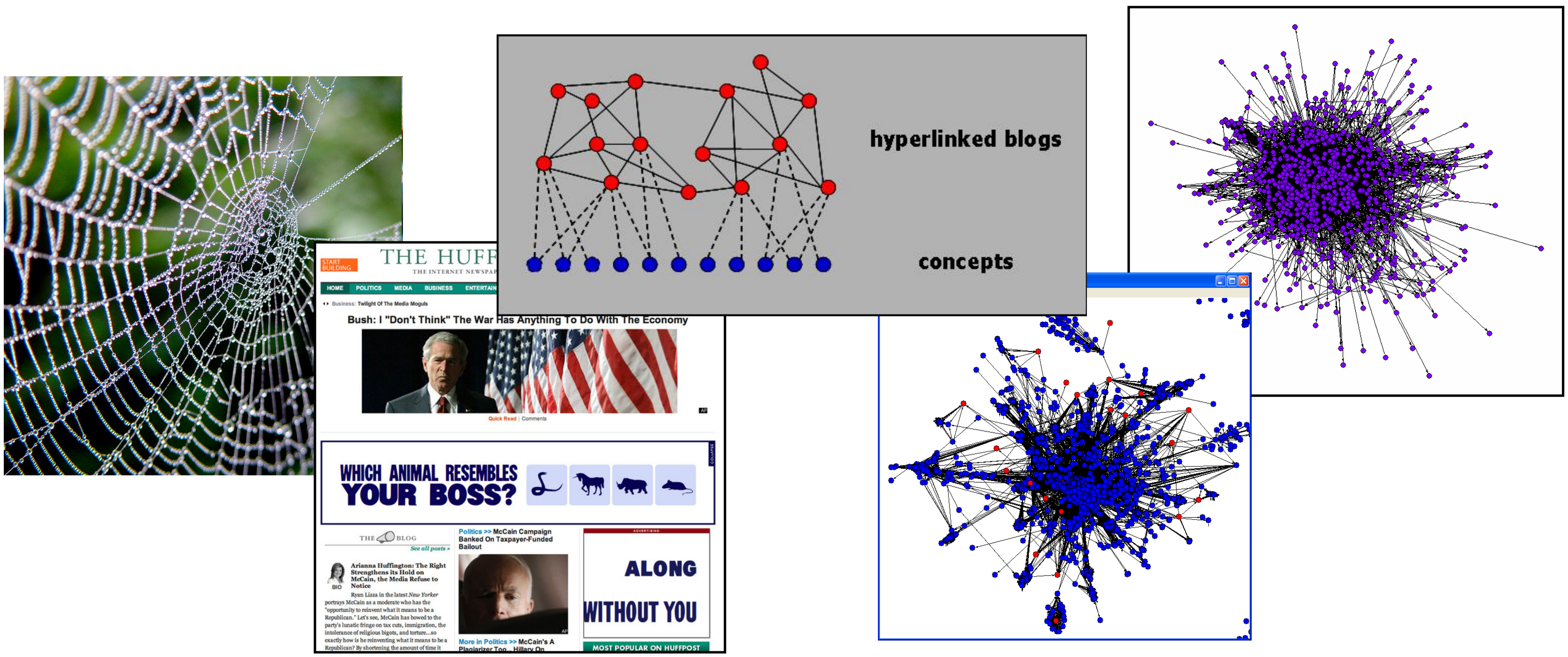
Predictive Analysis

Metric	Predictive?
post entropy	yes ($p<0.002$)
post volume	no
lexicon intrinsics	no

‘Known keywords’ alert: web crawl

Given set of keywords associated with an emerging (or potential) event of interest, we describe a four step process for generating warning alerts.

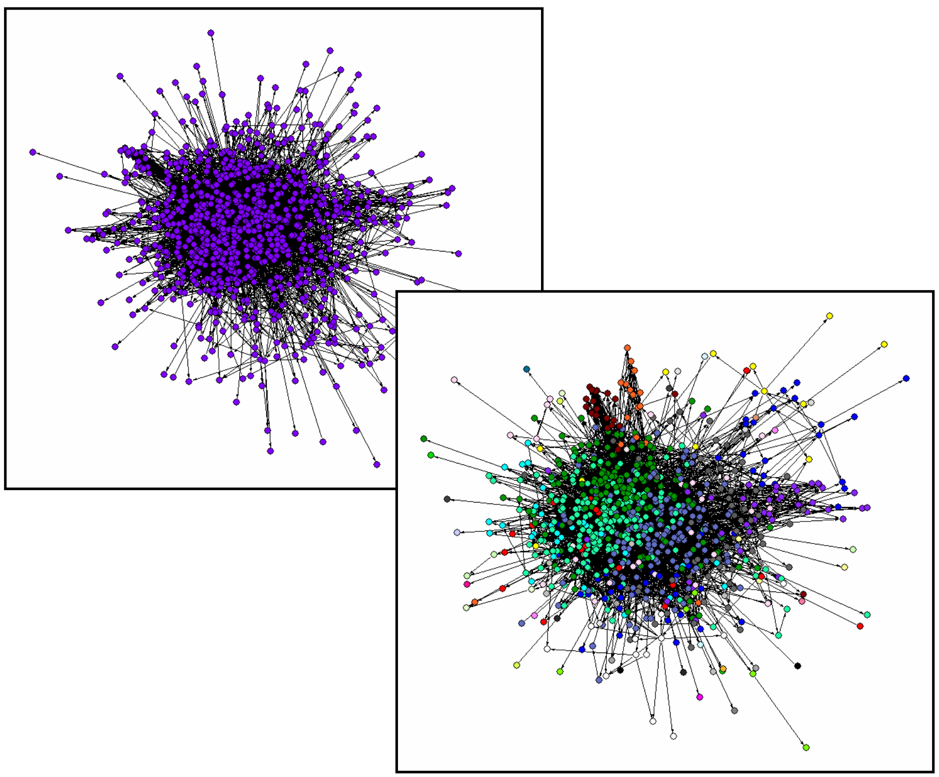
Step One: Perform web crawl, construct blog graph, detect keywords and timestamps in posts.

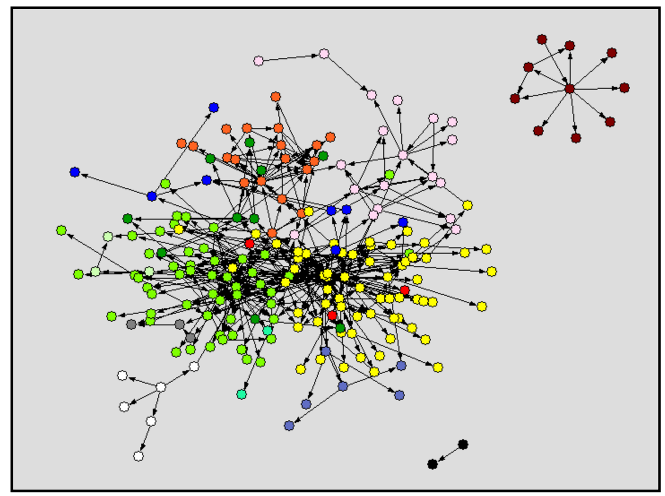


'Known keywords' alert: communities

'Known keywords' alerting procedure (cont'd)

Step Two: Partition blog graph into social network communities using new weighted CNM (wCNM) algorithm.

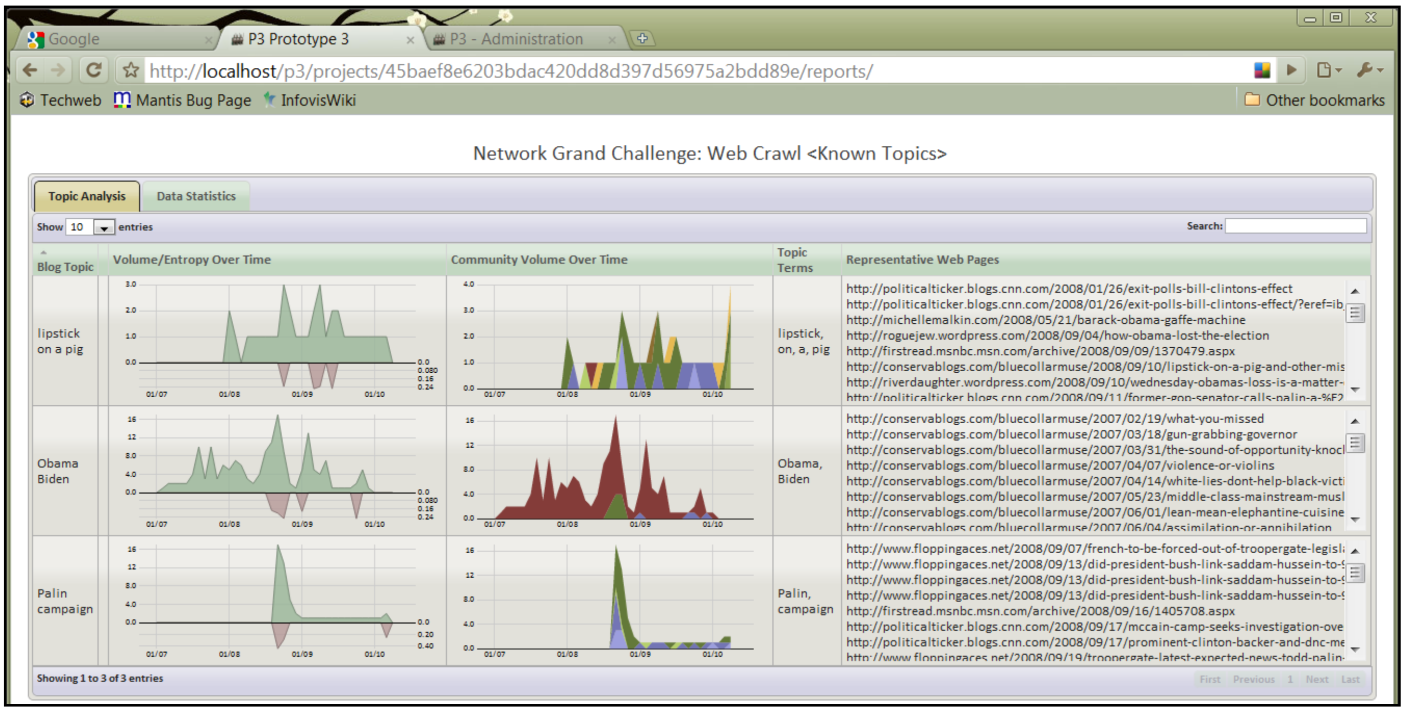




'Known keywords' alert: volume, entropy

'Known keywords' alerting procedure (cont'd)

Step Three: Assemble time series of post volume and post entropy



‘Known keywords’ alert: compare to ‘normal’

‘Known keywords’ alerting procedure (cont’d)

Step Four: Decide if post entropy ent(t) is ‘large’ and, if so, send alert.

- Method 1: multiplier method.

  Offline:

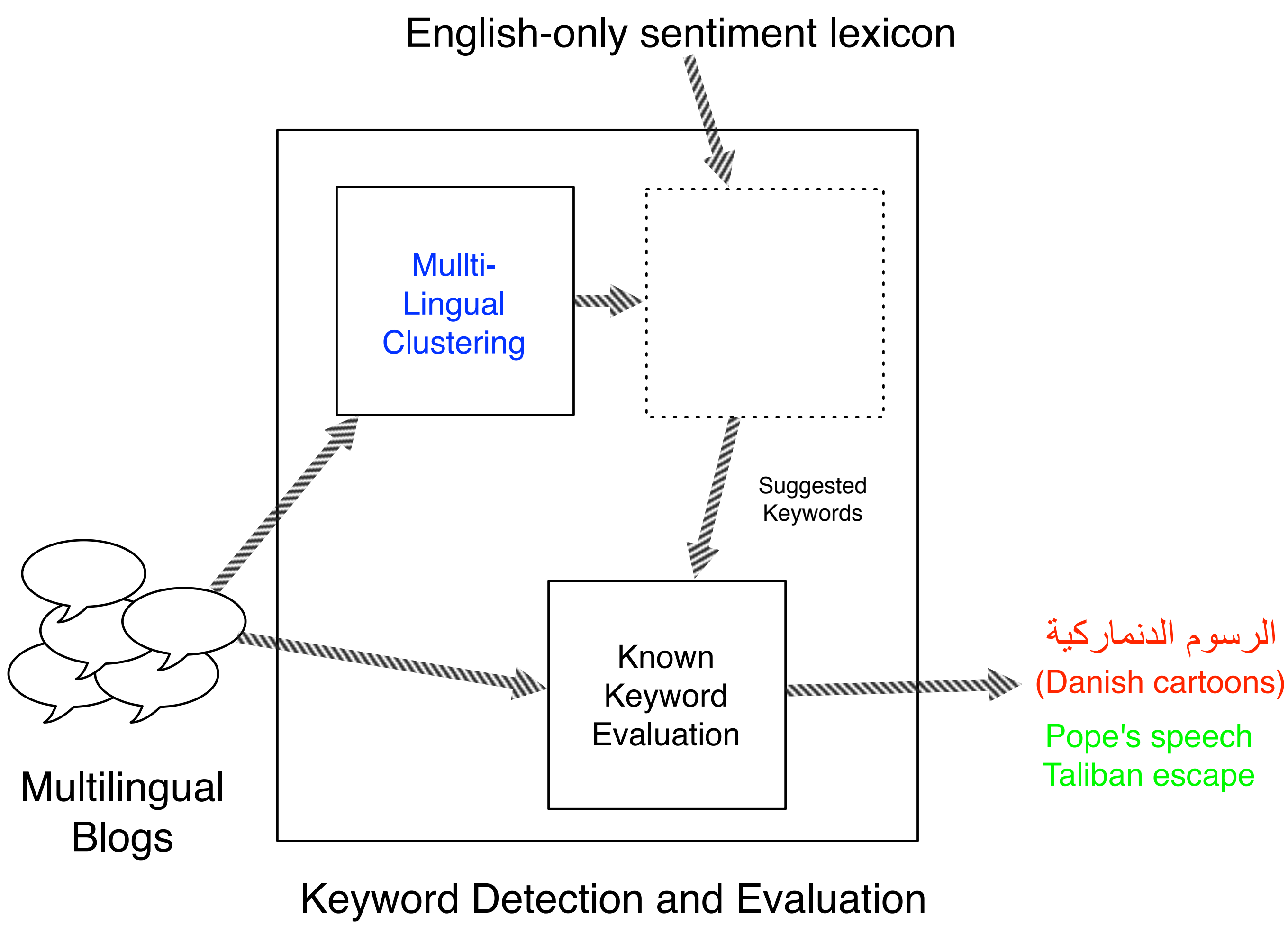
  1. Construct S-HDS social diffusion model that possesses correct social community graph and produces ‘stylized facts’ of vol(t) dynamics.
  2. Generate ensemble of entropy time series $\{sent_i(t)\}$ corresponding to null hypothesis: diffusion initiates in one (or a few) communities.
  3. Determine multiplier λ such that: $\lambda \times vol(t) \approx \mu\{sent_i(t)\}+\sigma\{sent_i(t)\}$.

  Online: Alert if $ent(t) > \lambda \times vol(t)$ for any t before ‘peak’ in vol(t).

- Method 2: direct method.

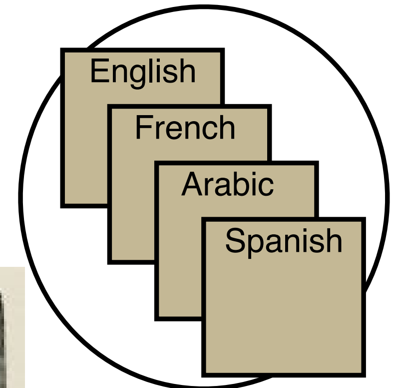
  Similar but construct $\{sent_i(t)\}$ online and compare ent(t) w/ $\{sent_i(t)\}$.

Multi-lingual Text Clustering



SNL has developed multilingual text analysis to link threats across multiple languages

- "Translate" new documents into a language-independent concept space, which is useful for:
  - Translation triage (i.e., translate documents in clusters of interest)
  - Ideological classification (e.g., hostile to U.S.)
  - Multilingual sentiment analysis



Sandia's database: 54 languages: >99% coverage of web

Afrikaans	Estonian	Norwegian
Albanian	Finnish	Persian (Farsi)
Amharic	French	Polish
Arabic	German	Portuguese
Aramaic	Greek (New Testament)	Romani
Armenian Eastern	Greek (Modern)	Romanian
Armenian Western	Hebrew (Old Testament)	Russian
Basque	Hebrew (Modern)	Scots Gaelic
Breton	Hungarian	Spanish
Chamorro	Indonesian	Swahili
Chinese (Simplified)	Italian	Swedish
Chinese (Traditional)	Japanese	Tagalog
Croatian	Korean	Thai
Czech	Latin	Turkish
Danish	Latvian	Ukrainian
Dutch	Lithuanian	Vietnamese
English	Manx Gaelic	Wolof
Esperanto	Maori	Xhosa

The Rosetta Stone

Bag of Words/Vector Space Model

example from (Berry, Drmac, Jessup, 1999)

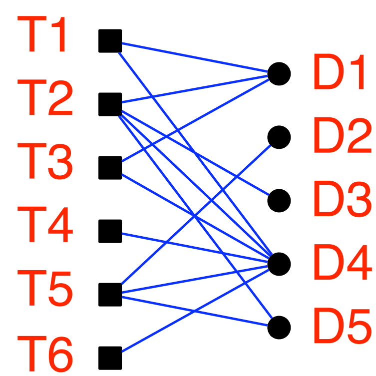
Documents

D1: How to Bake Bread Without Recipes
D2: The Classic Art of Viennese Pastry
D3: Numerical Recipes: The Art of Scientific Computing
D4: Breads, Pastries, Pies and Cakes: Quantity Baking Recipes
D5: Pastry: A Book of Best French Recipes

Terms

T1: bak(e,ing)
T2: recipes
T3: bread
T4: cake
T5: pastr(y,ies)
T6: pie

Bipartite graph



Term-by-doc (adjacency) matrix

$$\hat{A} = \begin{array}{c@{}c} \begin{array}{ccccc} D1 & D2 & D3 & D4 & D5 \end{array} & \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} & \begin{array}{c} T1 \\ T2 \\ T3 \\ T4 \\ T5 \\ T6 \end{array} \end{array}$$

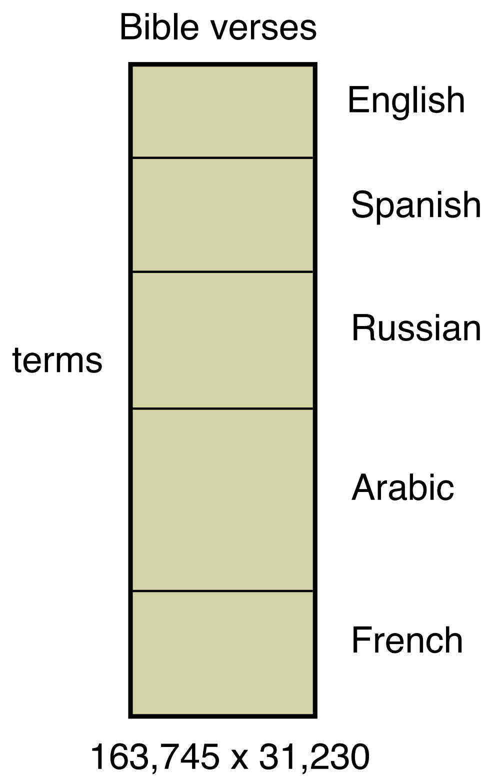
Key concepts

- Bag of words
- Stemming
- Vector space model
- Scaling for information content

Sandia National Laboratories

Term-Doc Matrix

Term-by-verse matrix
for all languages



Look for co-occurrence of terms in the same verses and across languages to capture latent concepts

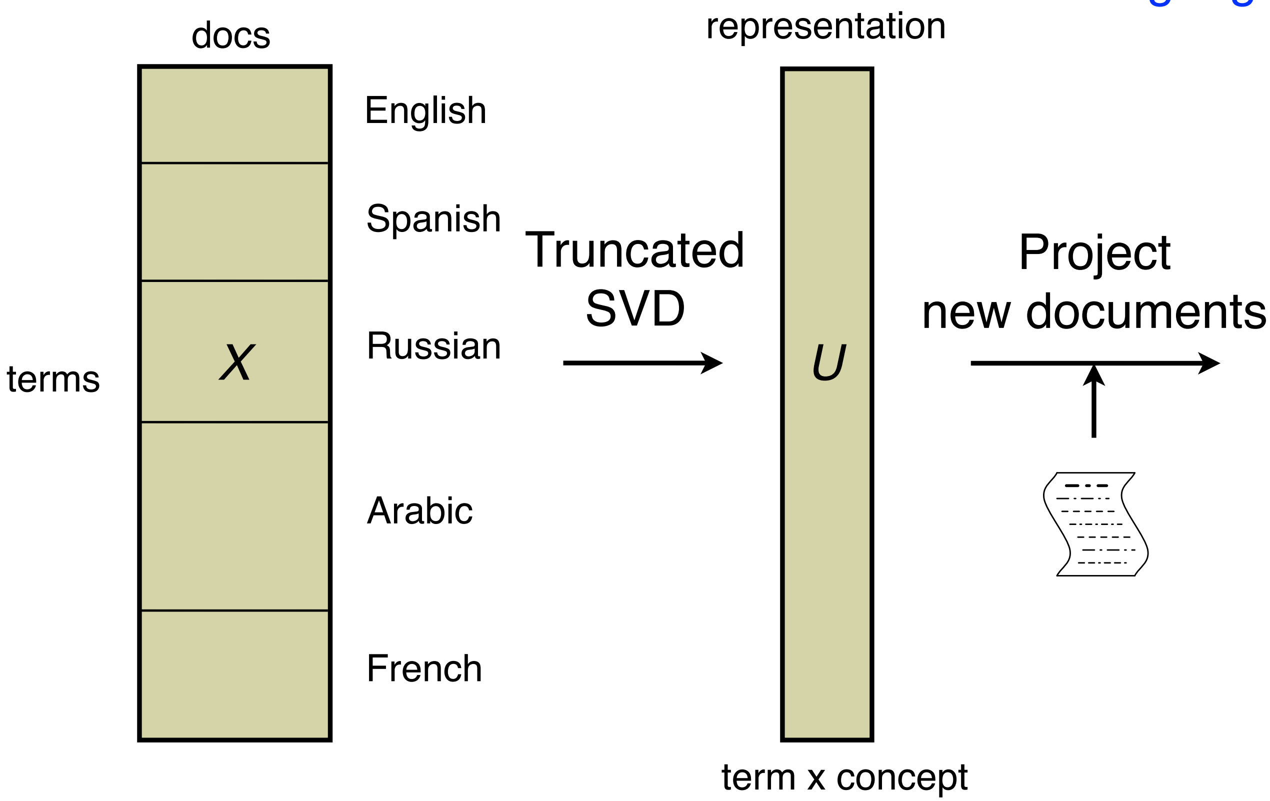
- Approach is not new: pairs of languages in Latent Semantic Analysis (LSA)
  - English and French (Landauer & Littman, 1990)
  - English and Greek (Young, 1994)
- *Multi-parallel* corpus is new

Europarl Corpus

- Extracted from the proceedings of the European Parliament
- Translations in 11 languages
  - French, Italian, Spanish, Portuguese (Romantic)
  - English, Dutch, German, Danish, Swedish (Germanic)
  - Greek
  - Finnish
- Sentence aligned text (16 M sentences across 11 languages)
- 1,247,832 speeches (including translations)
- 1,249,253 terms (from all 11 languages)
  - English terms: 46,074

Multilingual Latent Semantic Analysis

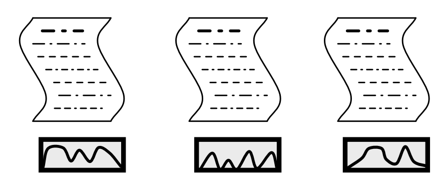
Term-by-doc matrix
for all languages



"Translate" new documents
into a small number of
language-independent features

dimension 1	0.1375
dimension 2	0.1052
dimension 3	0.0341
dimension 4	0.0441
dimension 5	-0.0087
dimension 6	0.0410
dimension 7	0.1011
dimension 8	0.0020
dimension 9	0.0518
dimension 10	0.0822
dimension 11	-0.0101
dimension 12	-0.1154
dimension 13	-0.0990
dimension 14	0.0228
dimension 15	-0.0520
dimension 16	0.1096
dimension 17	0.0294
dimension 18	0.0495
dimension 19	0.0553
dimension 20	0.1598

Document feature vector

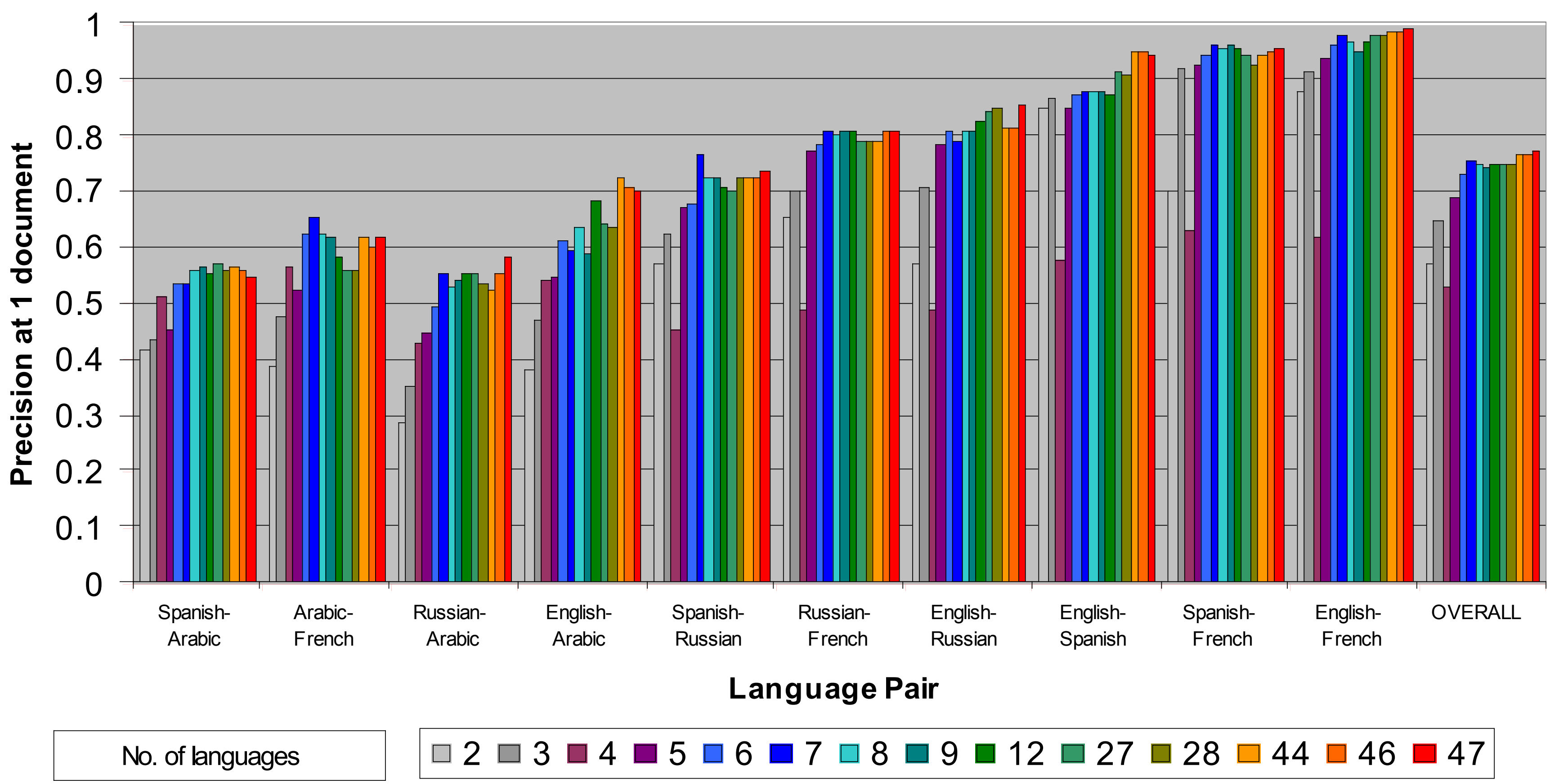


Applications
- cross-language retrieval
- pairwise similarities for clustering
- machine learning applications

More languages = Better results

(Chew and Abdelali, 2007)

LSA with 300 concept vectors



Language Morphology

Translation	Terms	Total Words
English (King James)	12,335	789,744
Arabic (Smith Van Dyke)	55,300	440,435

Languages convey information in different number of words

Isolating language	Synthetic language
Chinese	Quechua, Inuit (Eskimo)

- Isolating language: One morpheme per word
  - e.g., "He travelled by hovercraft on the sea." Largely isolating, but travelled and hovercraft each have two morphemes per word. (Wikipedia)
- Synthetic language: High morpheme-per-word ratio
  - German: *Aufsichtsratsmitgliederversammlung* => "On-view-council-with-limbs-gathering" meaning "meeting of members of the supervisory board". (Wikipedia)
  - Chulym: *Aalychtypiskem* => "I went out moose hunting"
  - Yup'ik Eskimo: *tuntussuqatarniksaitengqiggtuq* => "He had not yet said again that he was going to hunt reindeer." (Payne, 1997)

Sample Tokenization

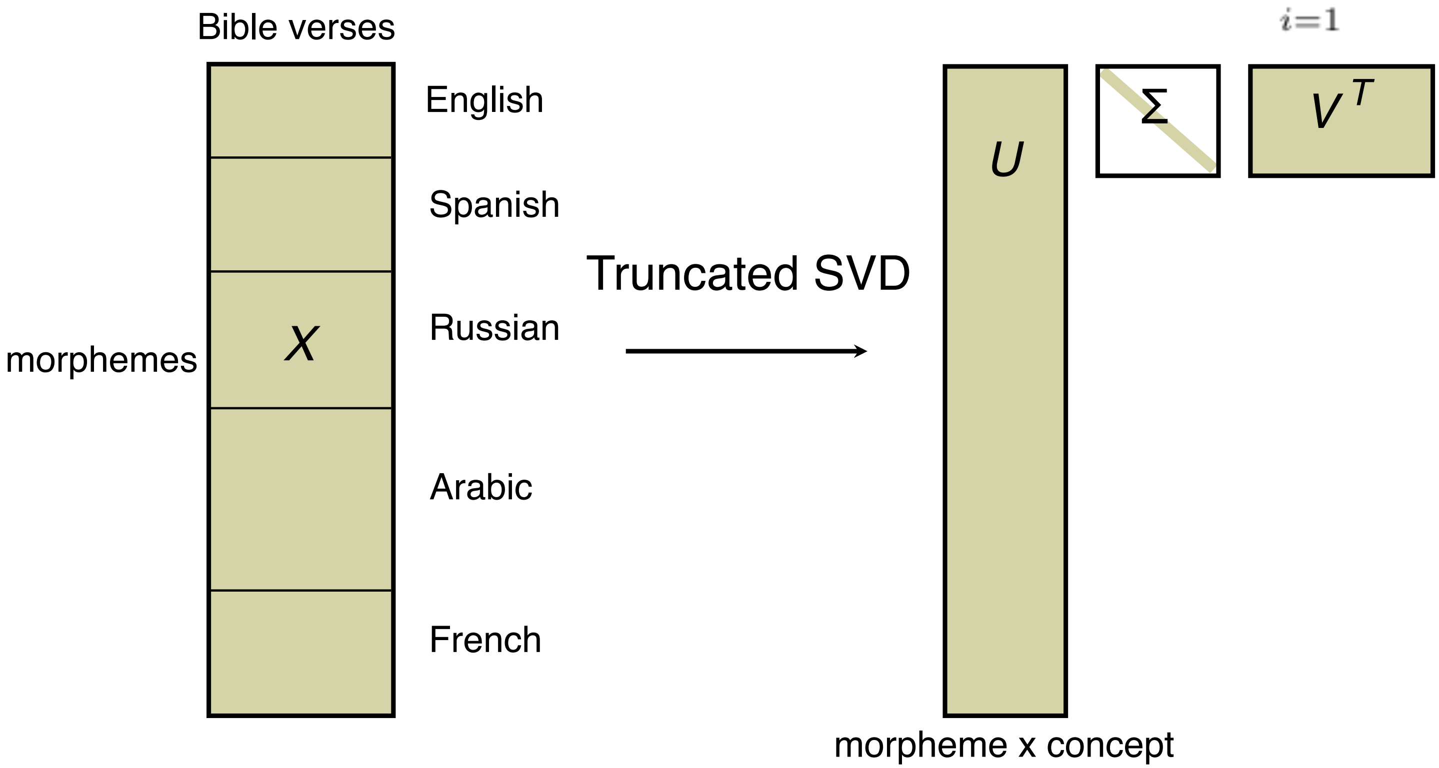
Wordform	Tokenization
abaissée	*abaissé + e*
abaissées	*abaissé + es*
abaissèrent	*abaiss + èrent*
acceptance	*accept + ance*
acceptation	*accept + ation*
acquaintance	*acquaint + ance*

We use these "morphemes"
in place of terms

Latent Morpho-Semantic Analysis (LMSA)

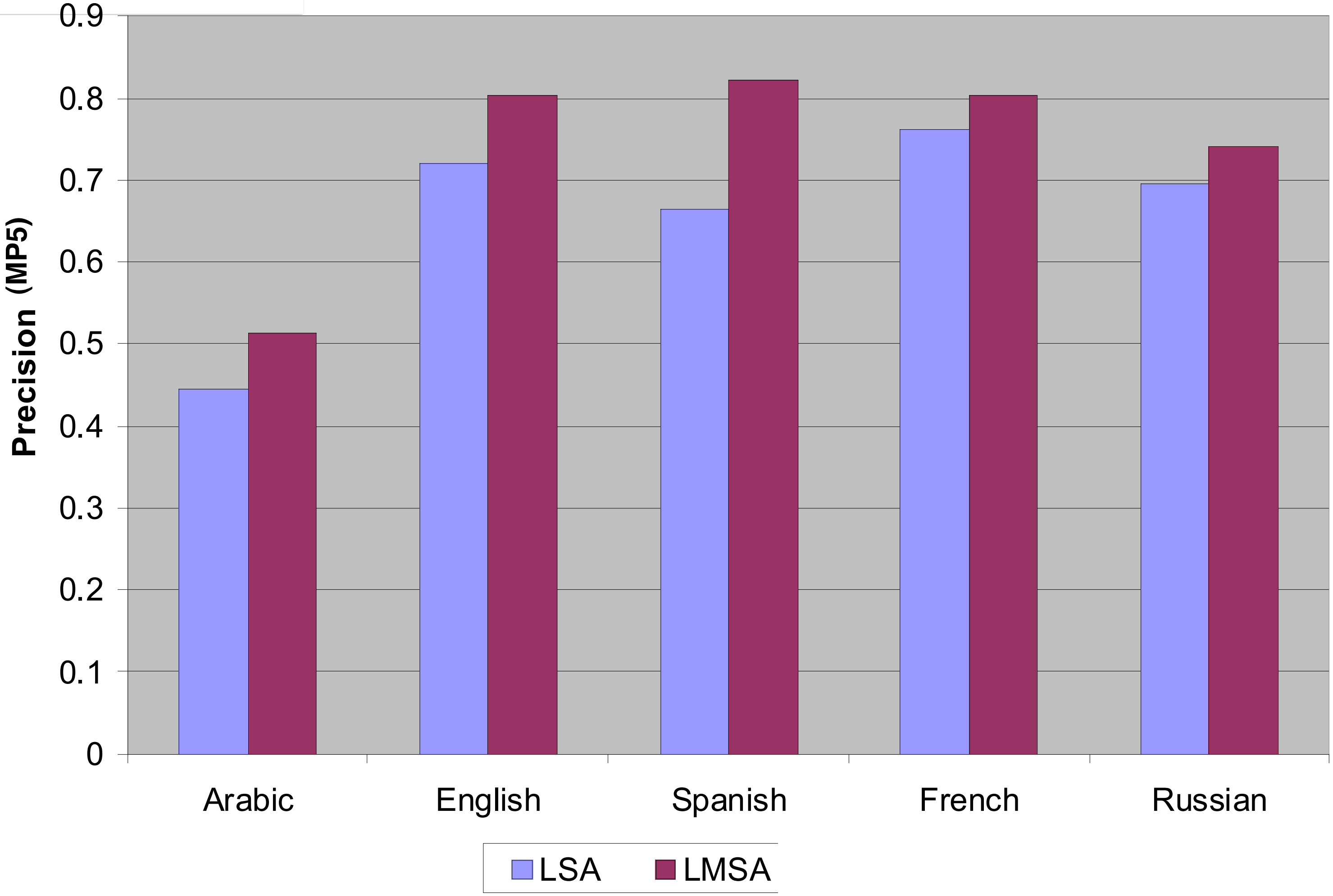
Morpheme-by-verse matrix
for all languages

$$X_{k} = U_{k}\Sigma_{k}V_{k}^{T} = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$



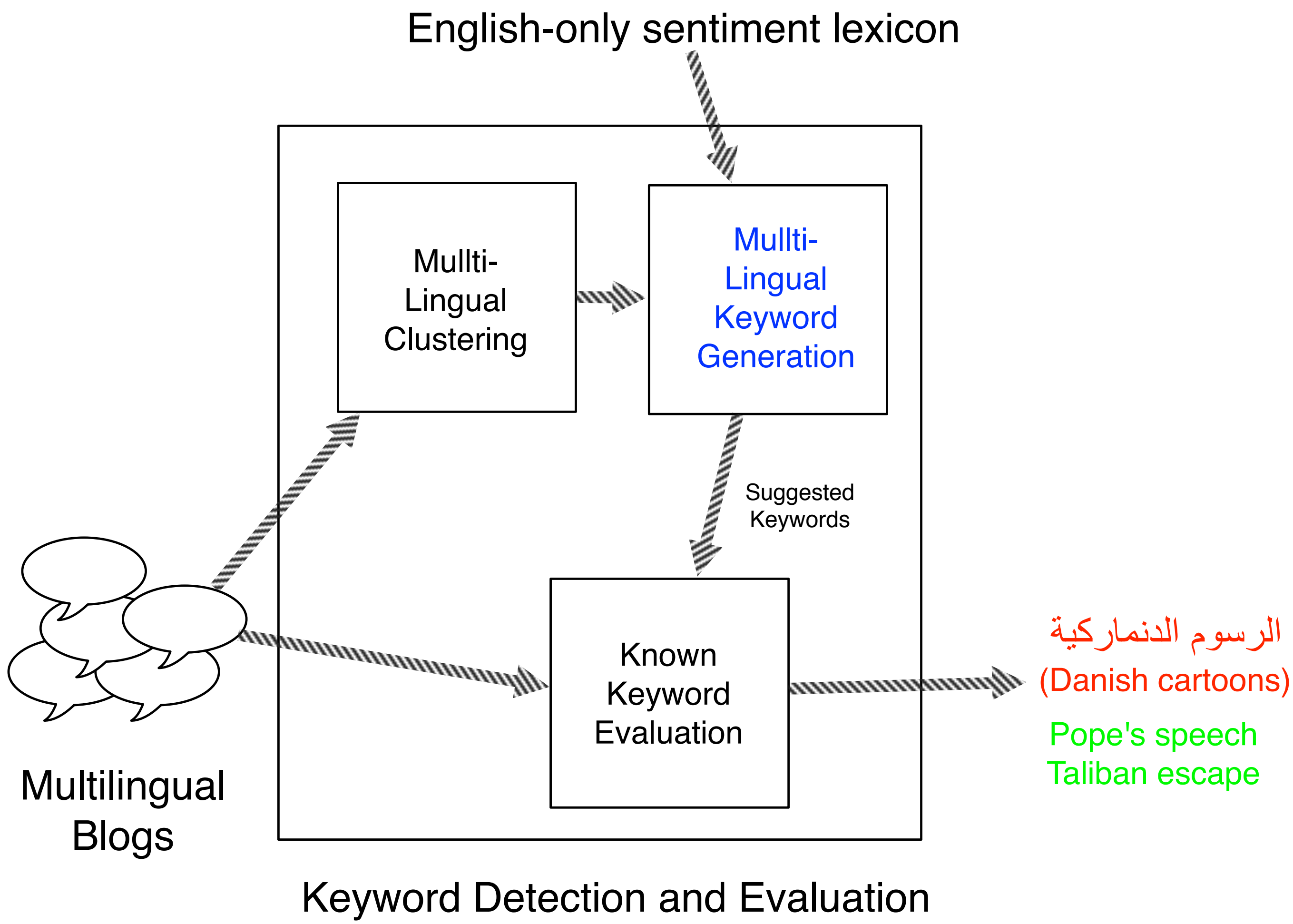
- Fewer morphemes than terms
- X matrix is smaller but denser

Comparison by Language

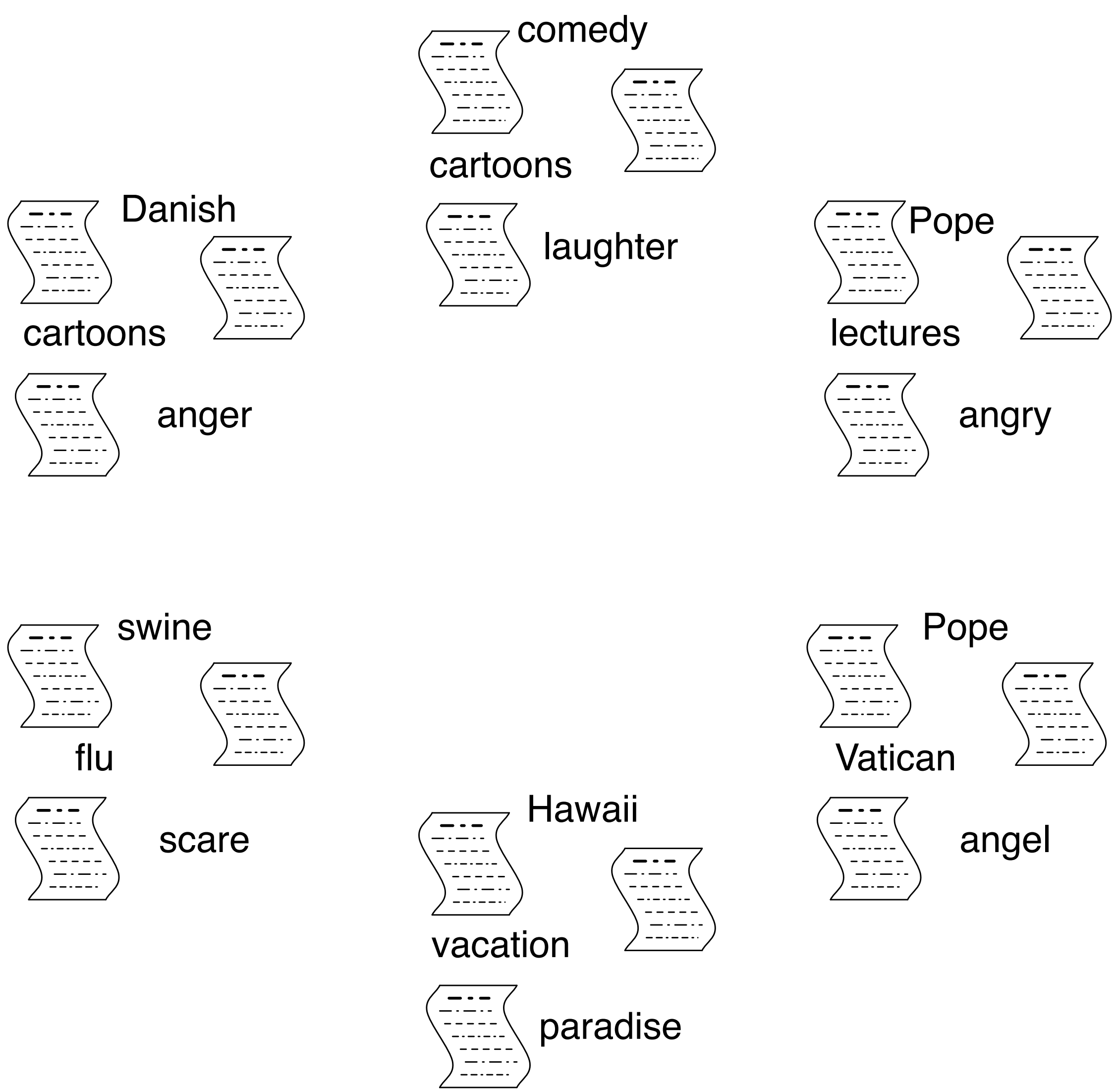


Statistically significant improvements at $p < 0.001$

Multi-lingual Sentiment Analysis

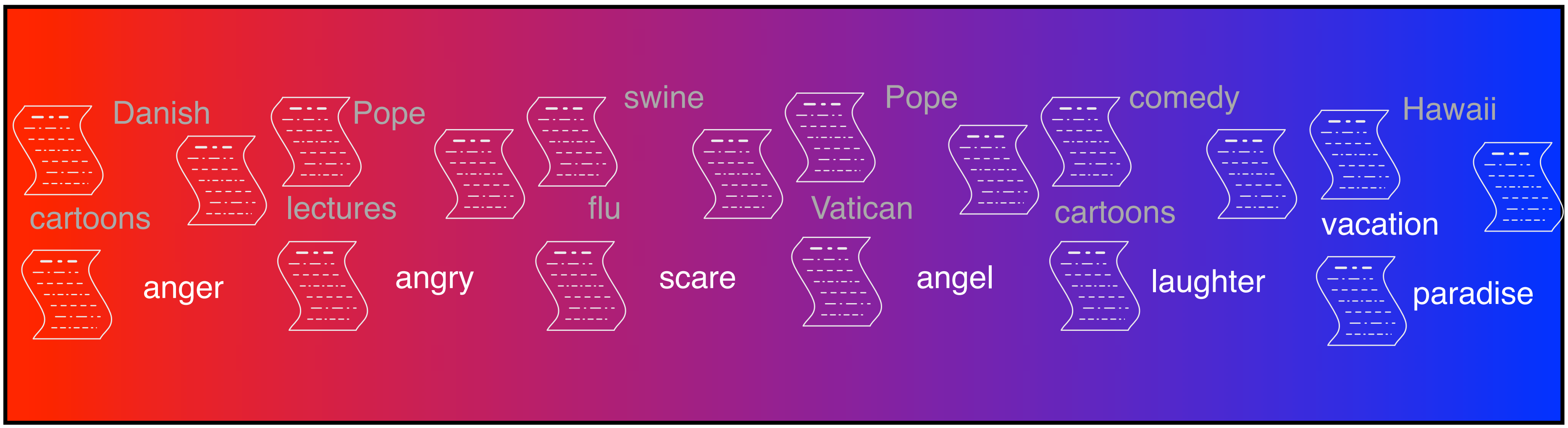


Conceptual Example



Conceptual Example

1) Sort the documents according to sentiment

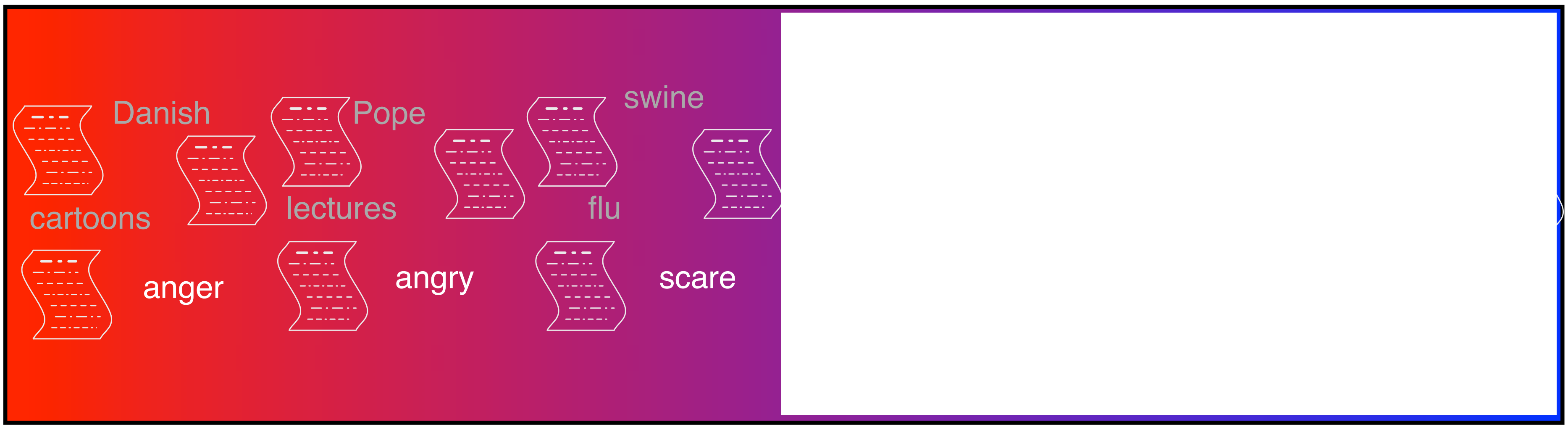


Unpleasant

Pleasant

Conceptual Example

2) Keep only the highly emotional documents

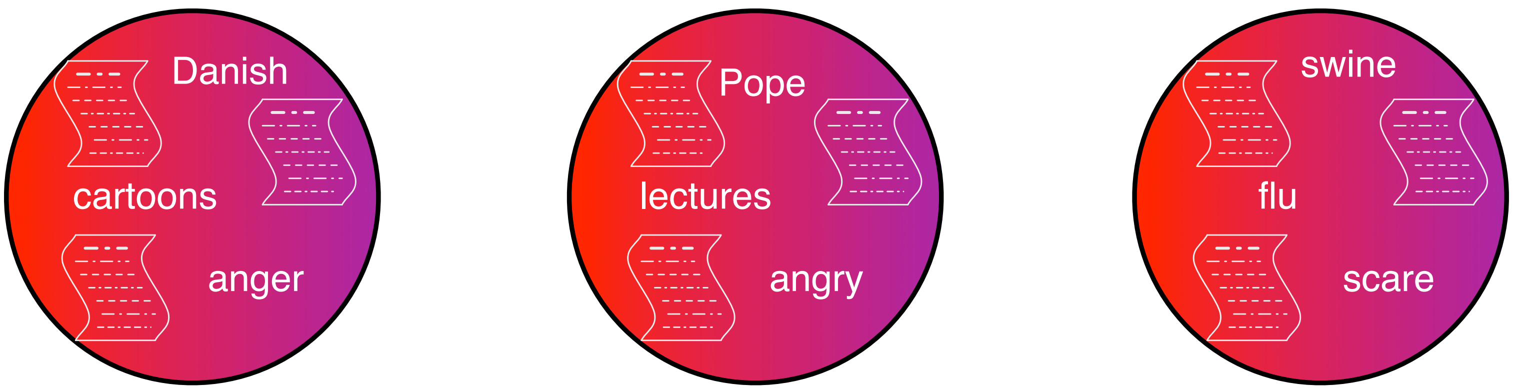


Unpleasant Ambivalent Pleasant

Conceptual Example

3) Cluster by topic using multilingual document clustering

4) Find unique keywords that describe each cluster



Danish, cartoons

Pope, lectures

swine, flu

Extracting Sentiment from Text

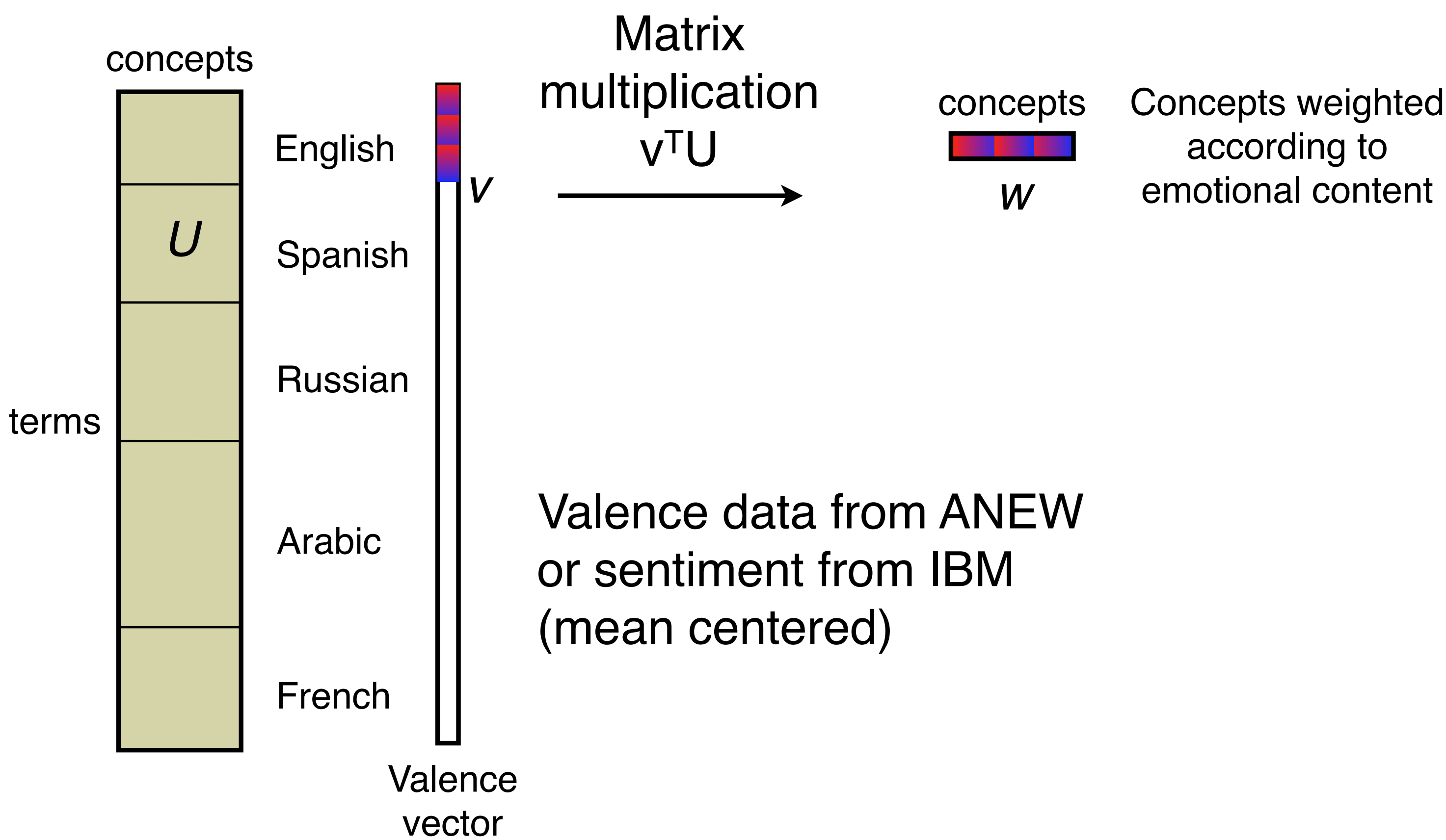
- In English, there are a number of sentiment lexicons and emotional valence dictionaries
  - IBM
  - Affective Norms for English Words (ANEW)
  - Harvard IV-4
  - Lasswell value dictionary
  - Whissell's Dictionary of Affect in Language (DAL)
  - ...

- We would like to avoid a dependence on sentiment lexicons for foreign languages
  - Otherwise complexity increases!
  - Need to keep up with evolving/changing language
  - IC interest in least-spoken languages whereas commercial systems target most-spoken languages
  - "Early warning sentiment analysis"

Sandia National Laboratories

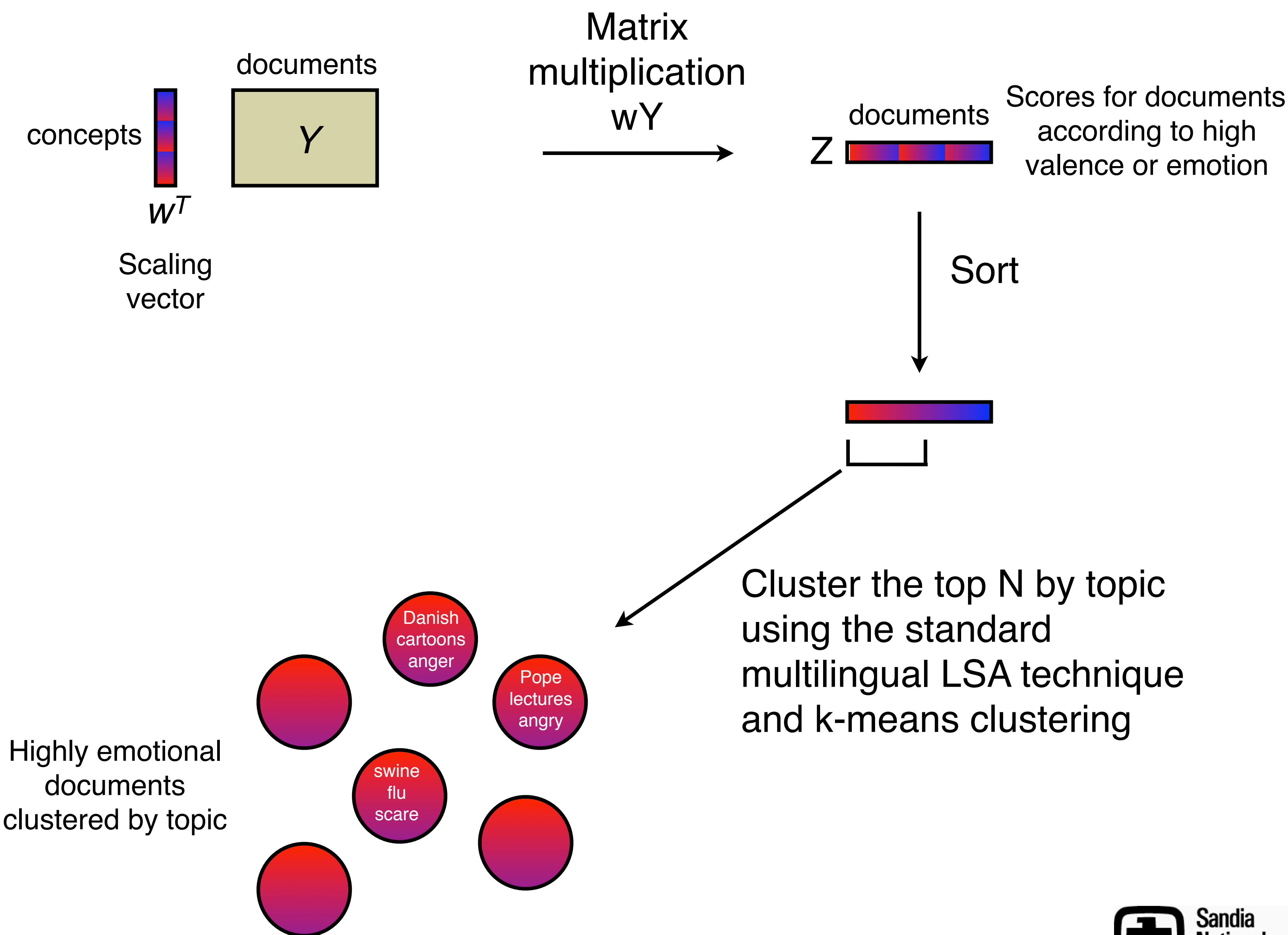
Concept Weighting and Scaling

Identify concepts associated with words with high valence

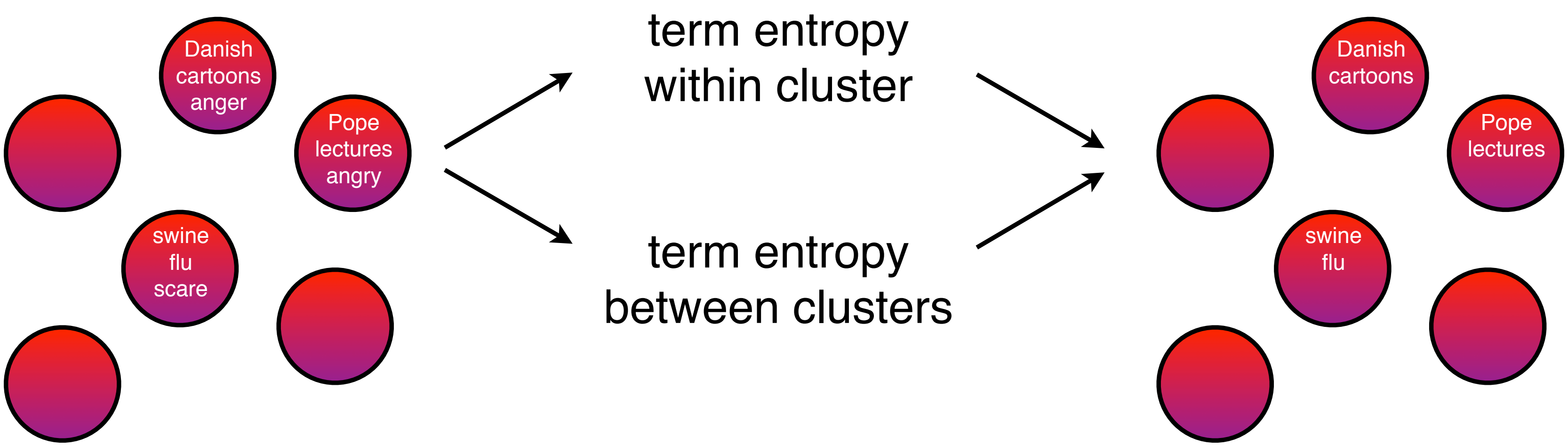


Valence data from ANEW or sentiment from IBM (mean centered)

Ranking, Sorting, Clustering



Identify Keywords

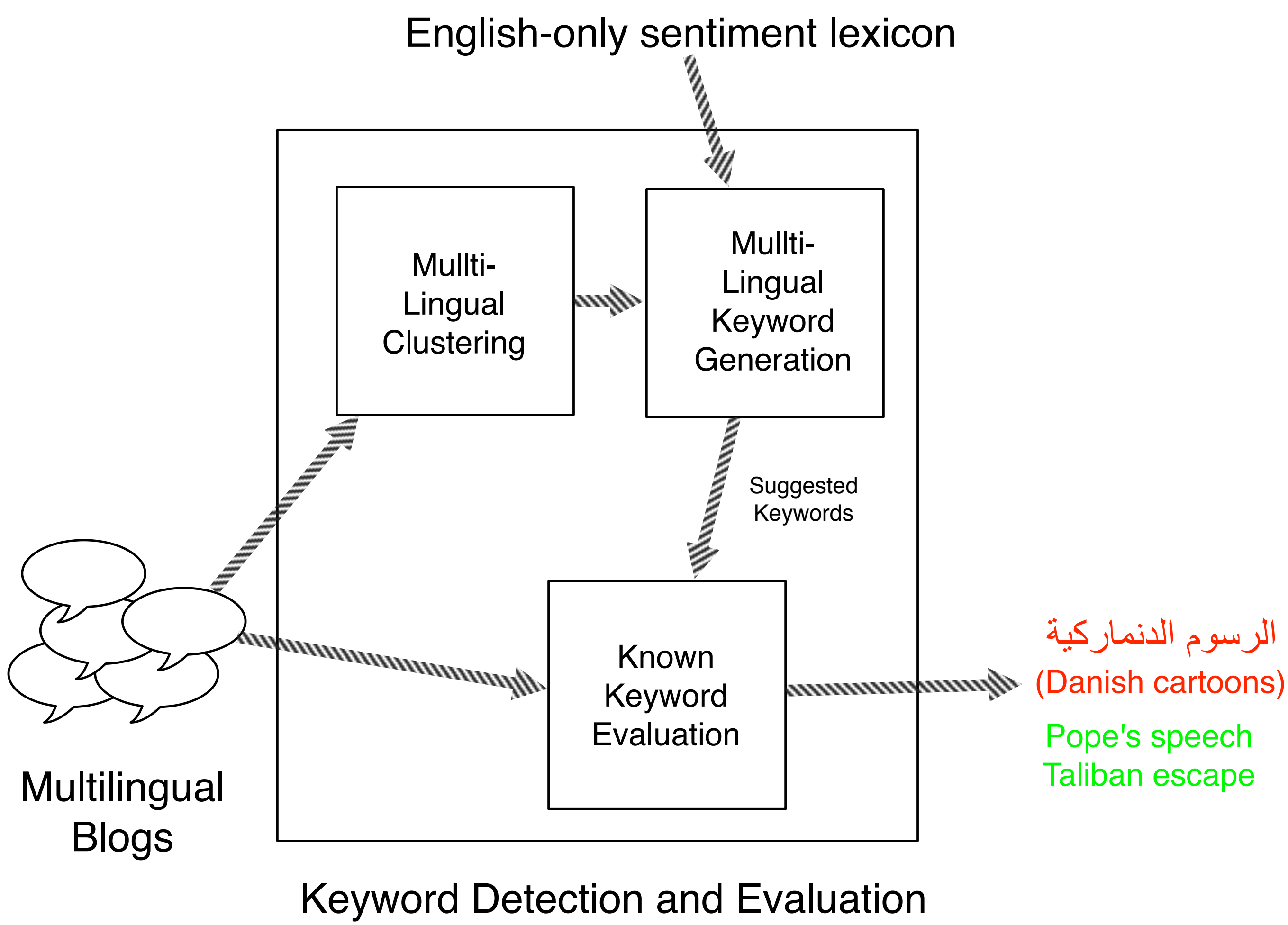


term counts in documents

✓	cartoon	5	4	3	0	0	0	1	0	0
	anger	2	3	1	1	3	1	0	1	2

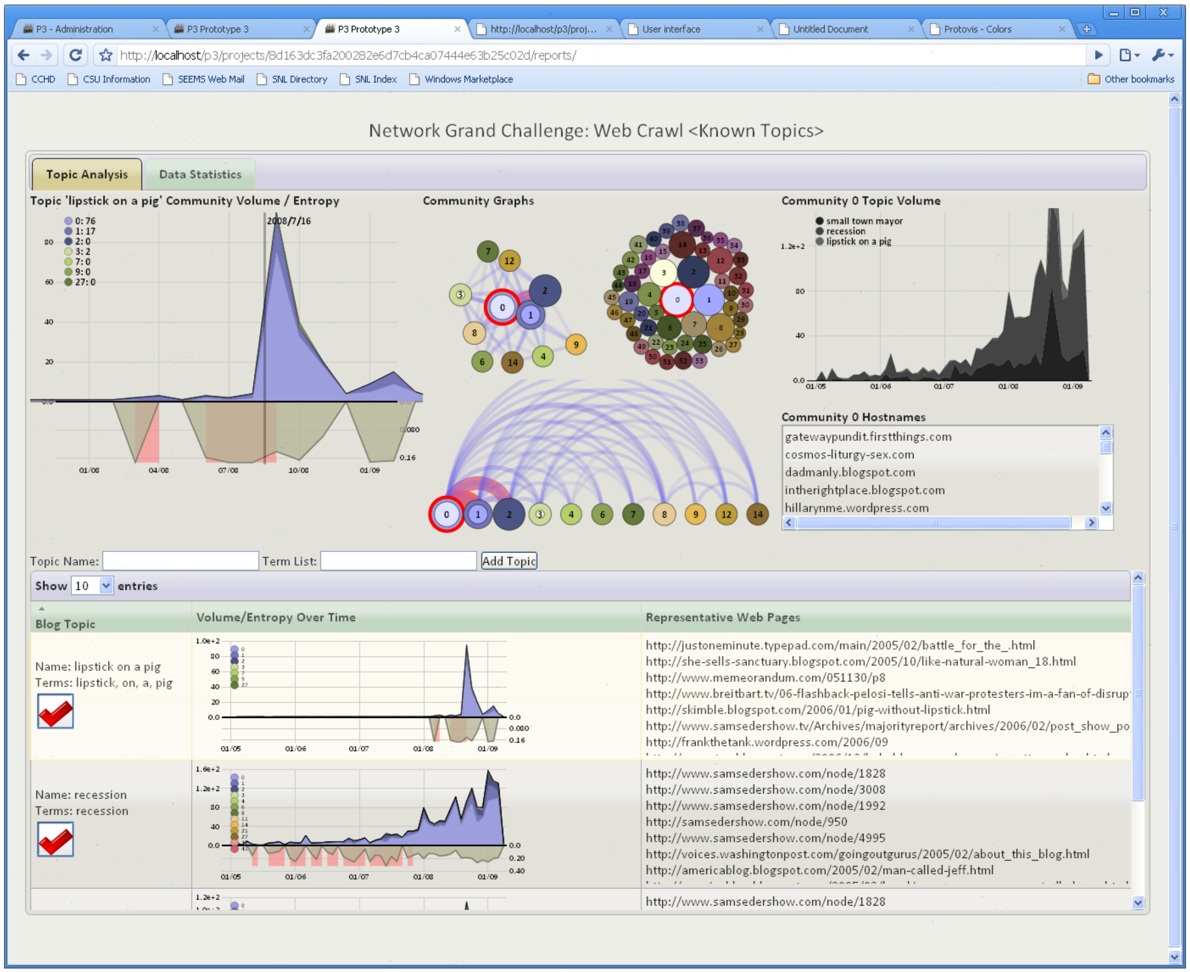
- Find unique terms that best describe each cluster.
- The idea is that we want to identify terms that have broad coverage within a cluster but don’t appear much outside of this cluster.
- Choose terms with low inter-cluster entropy but high intra-cluster entropy (e.g., by dividing the intra- by inter-entropy scores and choosing the top N terms per cluster on this scale).

The Black Box, Unpacked . . .



All Implemented in an Analyst's Tool

Interactive web interface, built in Titan and ProtoVis



CEP: Questions To Address (Part 1)

- What is our definition of CEP?
- **How do we do the processing?**

  Nightly web crawls, multi-lingual clustering, keyword generation and evaluation.
- **What kind of results do we get?**
  - Red/Green binary keyword alerts.
  - Volume and entropy time series plots for all keywords.
  - Drill down to source blogs at any point.

CEP: Questions To Address (Part 2)

- **Who is using our service and for what?**

  Sandia analysts, for cultural reaction assessments.

  Capabilities shared with DoD, IC, JASONS.

- **What are the computing and data specifications, limitations, metrics?**
  - Primary limitations: pairwise similarity calculations, ugliness of web data.
  - Metrics: customer anecdotes. Clustering validation, keyword replication, and sentiment prediction studies. Peer-reviewed publication.

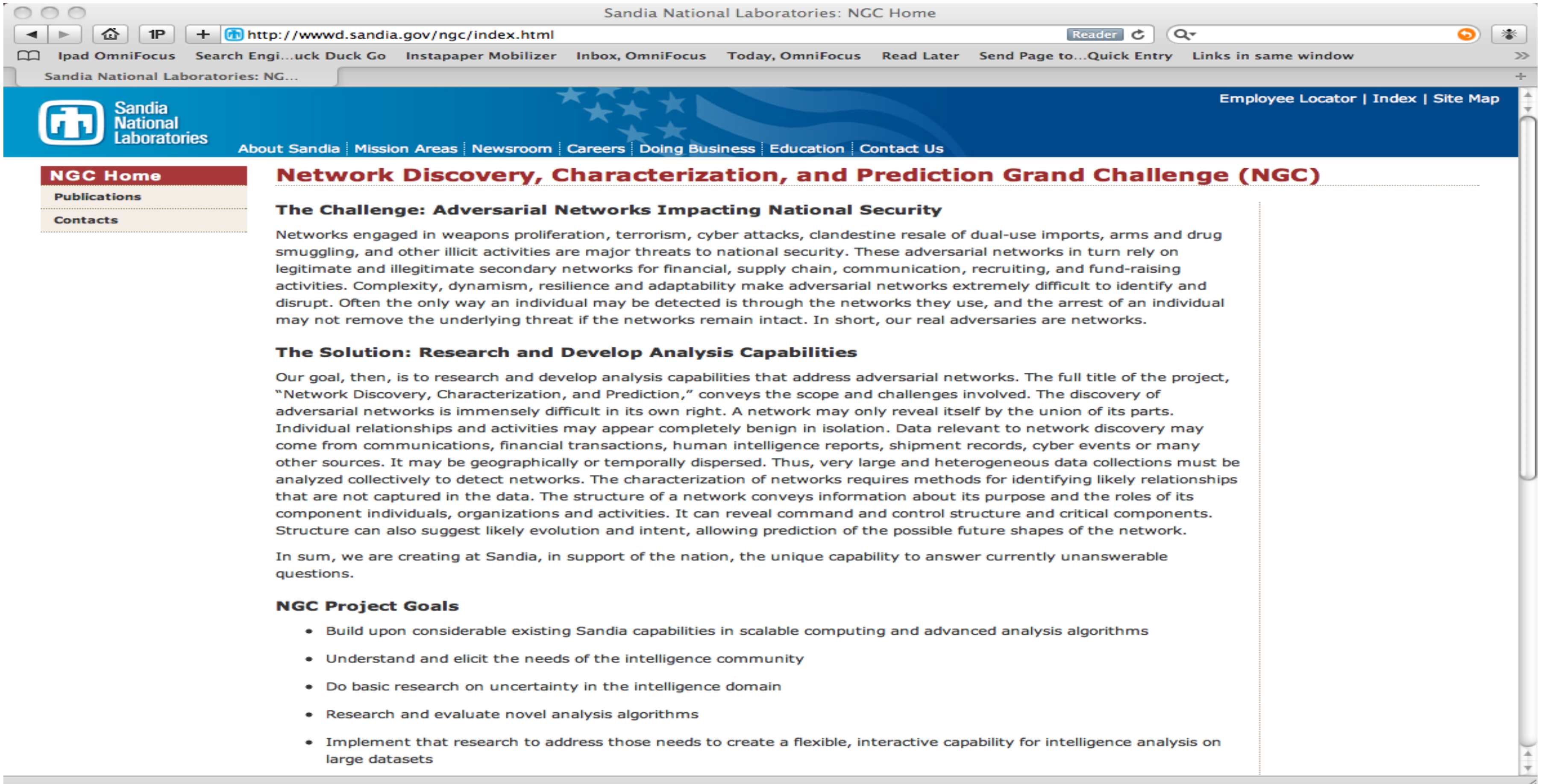
- **What examples of the whole process and success stories?**

  Bali-bomber execution in early November 2008; correctly predicted that outrage was **not** self-sustaining.

  Israel/Gaza conflict in early January 2009; correctly predicted that hacking discussion **was** self-sustaining, then actual hacking occurs.

For More Information . . .

- NGC Final Report, all publications, all contact info: `ngc.sandia.gov`



Sandia National Laboratories: NGC Home

http://wwwd.sandia.gov/ngc/index.html

Ipad OmniFocus | Search Engi...uck Duck Go | Instapaper Mobilizer | Inbox, OmniFocus | Today, OmniFocus | Read Later | Send Page to...Quick Entry | Links in same window

Sandia National Laboratories: NG...

Sandia National Laboratories

Employee Locator | Index | Site Map

About Sandia | Mission Areas | Newsroom | Careers | Doing Business | Education | Contact Us

NGC Home

Publications

Contacts

Network Discovery, Characterization, and Prediction Grand Challenge (NGC)

The Challenge: Adversarial Networks Impacting National Security

Networks engaged in weapons proliferation, terrorism, cyber attacks, clandestine resale of dual-use imports, arms and drug smuggling, and other illicit activities are major threats to national security. These adversarial networks in turn rely on legitimate and illegitimate secondary networks for financial, supply chain, communication, recruiting, and fund-raising activities. Complexity, dynamism, resilience and adaptability make adversarial networks extremely difficult to identify and disrupt. Often the only way an individual may be detected is through the networks they use, and the arrest of an individual may not remove the underlying threat if the networks remain intact. In short, our real adversaries are networks.

The Solution: Research and Develop Analysis Capabilities

Our goal, then, is to research and develop analysis capabilities that address adversarial networks. The full title of the project, "Network Discovery, Characterization, and Prediction," conveys the scope and challenges involved. The discovery of adversarial networks is immensely difficult in its own right. A network may only reveal itself by the union of its parts. Individual relationships and activities may appear completely benign in isolation. Data relevant to network discovery may come from communications, financial transactions, human intelligence reports, shipment records, cyber events or many other sources. It may be geographically or temporally dispersed. Thus, very large and heterogeneous data collections must be analyzed collectively to detect networks. The characterization of networks requires methods for identifying likely relationships that are not captured in the data. The structure of a network conveys information about its purpose and the roles of its component individuals, organizations and activities. It can reveal command and control structure and critical components. Structure can also suggest likely evolution and intent, allowing prediction of the possible future shapes of the network.

In sum, we are creating at Sandia, in support of the nation, the unique capability to answer currently unanswerable questions.

NGC Project Goals

- Build upon considerable existing Sandia capabilities in scalable computing and advanced analysis algorithms
- Understand and elicit the needs of the intelligence community
- Do basic research on uncertainty in the intelligence domain
- Research and evaluate novel analysis algorithms
- Implement that research to address those needs to create a flexible, interactive capability for intelligence analysis on large datasets

- Predictions from communities: Rich Colbaugh, `rcolbau@sandia.gov`, 505 284-4116
- Multilingual text analysis: Brett Bader, `bwbader@sandia.gov`, 505 845-0514
- PI, and sentiment analysis: Philip Kegelmeyer, `wpk@sandia.gov`, 925 294-3016